Scientific Journals
An international open-access platform

Hanlin Letters

Article

OPEN ACCESS

Multimodal Prediction of Early Clinical Deterioration in Adult Intensive Care Units: A Multi-Center Study

Wei Zhang,[1] Jian Liu,[2] Hao Chen[3]

[1] School of Computer Science and Technology, Hubei University of Technology, Nanli Road 28, Hongshan District, Wuhan 430068, China;
[2] School of Information Engineering, Zhejiang University of Science and Technology, Liuhe Road 318, Xihu District, Hangzhou 310023, China;
[3] College of Computer and Information Engineering, Henan Normal University, Jianshe East Road 46, Xinxiang 453007, China;

Abstract

Large intensive care units generate dense streams of physiological observations, laboratory values, orders, and narrative documentation, yet the translation of these heterogeneous data into stable early warning signals remains uneven across hospitals. In China, this challenge is intensified by cross-hospital variation in documentation practice, measurement frequency, and information system architecture. We present an empirical study of early deterioration prediction using a multimodal learning framework designed for adult intensive care units in a multi-center Chinese setting. The study uses de-identified electronic health records from six tertiary hospitals spanning northern, eastern, and southwestern China, with hourly prediction windows constructed from vital signs, laboratory trajectories, medication and device indicators, and Mandarin clinical notes. The target outcome is a composite deterioration event within the next 24 hours, defined as ICU death, initiation of invasive mechanical ventilation, vasopressor initiation, or emergency continuous renal replacement therapy. The proposed model combines a missingness-aware temporal transformer for structured signals, a domain-adapted Chinese clinical language encoder for free text, and a hospital-invariant fusion module trained to preserve predictive content while reducing site-specific bias. Across 5,931,284 prediction windows from 118,406 ICU stays, the multimodal model achieves an internal test AUROC of 0.867 and AUPRC of 0.352, outperforming structured-only, text-only, logistic regression, gradient boosting, and recurrent baselines. On a fully held-out external hospital, performance remains stable with AUROC 0.846 and AUPRC 0.327. Calibration improves materially after temperature scaling, and operational simulation shows that ranking by model score can concentrate a substantial fraction of future deterioration events within a manageable alert budget. The findings support multimodal, cross-hospital, and calibration-aware modeling as practical directions for ICU risk estimation in Chinese critical care systems.

1. Introduction

Early deterioration prediction in the intensive care unit is a systems problem as much as a modeling problem [1]. The ICU is saturated with measurements, but the raw availability of data does not guarantee timely recognition of impending instability. Risk often accumulates through weakly aligned signals: a rising lactate that precedes vasopressor initiation by several hours, a subtle reduction in urine output that is not decisive in isolation, a nurse note recording worsening agitation, or a physician assessment that suggests concern before hard thresholds are crossed. Conventional rule-based monitoring captures only a narrow projection of this state space. In high-throughput Chinese tertiary hospitals, where patient load is large and workflow heterogeneity across departments is common, the limits of threshold-triggered surveillance become more pronounced. Different units measure different variables at different cadences, clinicians document using varied note styles, and locally optimized practices can produce distribution shifts that undermine naive single-site models. An effective deterioration model in this environment must therefore address four simultaneous requirements: it must preserve fine temporal structure, exploit textual context, tolerate informative missingness, and generalize across hospitals whose electronic health record ecosystems are only partially standardized.

Recent machine learning work has shown that structured ICU trajectories contain predictive information well before overt decompensation becomes obvious to bedside observation. Yet structured measurements are not the whole record. Clinicians routinely encode anticipatory judgments in free text, and those judgments often summarize latent context that is difficult to reconstruct from numeric variables alone. The immediate motivation for integrating text and physiologic trajectories comes from evidence that note fusion can improve deterioration prediction over structured-only modeling; Sadanandan (2025) reported a 2.5 percentage-point AUROC gain and a 39.2% relative AUPRC gain when clinical notes were added to ICU time-series predictors [2]. This is a useful signal for model design because it suggests that unstructured clinical language can contribute complementary information rather than merely echo existing charted measurements [3]. In the Chinese ICU context, however, multimodal modeling faces additional engineering difficulties. Clinical notes are shorter and more formulaic in some hospitals, more discursive in others, and often include unit abbreviations, transliterated drug names, department-specific shorthand, and embedded numerals that complicate standard tokenization. A model that performs well on English ICU benchmarks cannot simply be ported without adaptation.

The present study formulates early deterioration prediction as a cross-hospital multimodal representation learning task. Instead of treating each hospital as an interchangeable sample source, we explicitly model the hospital dimension as a source of structured variation that should be attenuated in the latent representation while preserving outcome-relevant content. This framing is motivated by routine practical observations in Chinese healthcare informatics. The same laboratory test may appear with different naming conventions. Documentation density differs between medical and surgical ICUs. Orders and devices may be recorded with site-specific templates. These differences are not incidental noise; they can be strong enough to let a model learn hospital identity more easily than patient risk. When that happens, apparent performance on mixed random splits can mask vulnerability to site shift [4]. A clinically usable system must therefore be optimized not only for discrimination but also for transportability, which in practice means that it should preserve performance when deployed at a previously unseen hospital with a different local charting style.

To address these issues, we design a model centered on three principles. The first is missingness-aware temporal encoding. ICU data are irregularly sampled, and the pattern of what is not measured can itself be predictive. Rather than discarding this signal through heavy imputation alone, we expose the model to observation masks and elapsed-time channels so that the representation can condition on both value and measurement process. The second principle is modality complementarity. Structured trajectories and narrative notes carry different kinds of information. Numeric sequences capture physiological dynamics, while notes encode causal hypotheses, treatment intent, and qualitative assessment. We therefore use separate encoders for the two modalities and fuse them through gated cross-attention rather than simple concatenation. The third principle is hospital-invariant learning. We train a site-adversarial component that penalizes representations that are overly predictive of the source hospital, thereby reducing the risk that performance depends excessively on institution-specific documentation artifacts [5].

The study is anchored in a large de-identified cohort assembled from six tertiary hospitals in China. The task is defined at hourly resolution beginning six hours after ICU admission. For each hour, the model observes the preceding 48 hours of structured data and the most recent available narrative context, then estimates the probability of a composite deterioration event within the next 24 hours. The composite endpoint includes ICU death, invasive mechanical ventilation initiation, vasopressor initiation, and emergency continuous renal replacement therapy. This definition intentionally combines events that represent clinically meaningful escalation of organ support, because the operational goal is not merely to classify mortality but to detect a broader transition from relative stability to resource-intensive instability. In a real ICU workflow, these events occupy overlapping decision pathways: hemodynamic failure, respiratory decline, renal decompensation, and terminal deterioration often emerge through interacting processes rather than cleanly separable labels.

The empirical contribution of this manuscript is not restricted to a new architecture. We also examine whether cross-hospital adaptation materially changes external performance, whether textual information remains useful after strong structured baselines are established, how calibration behaves before and after post-hoc correction, and how alert concentration changes under practical workload constraints. These questions are central to computer science for health settings because discrimination alone is a narrow metric. A model can have a favorable AUROC yet remain poorly calibrated, unstable under site shift, or operationally unusable when converted into alerts. We therefore evaluate the proposed method against logistic

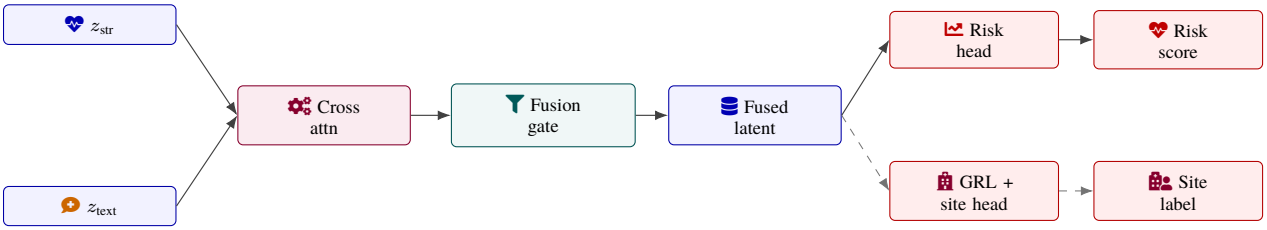

Figure 1: Fusion and hospital-robust training. Structured and textual states first exchange information through cross-modal attention, then pass through a gate that adjusts modality contribution at each patient-hour. The fused latent state feeds the clinical risk head, while a gradient-reversed site head discourages hospital-specific shortcuts. This diagram captures the manuscript's transportability argument: preserve predictive content, but reduce representation entanglement with local charting style.

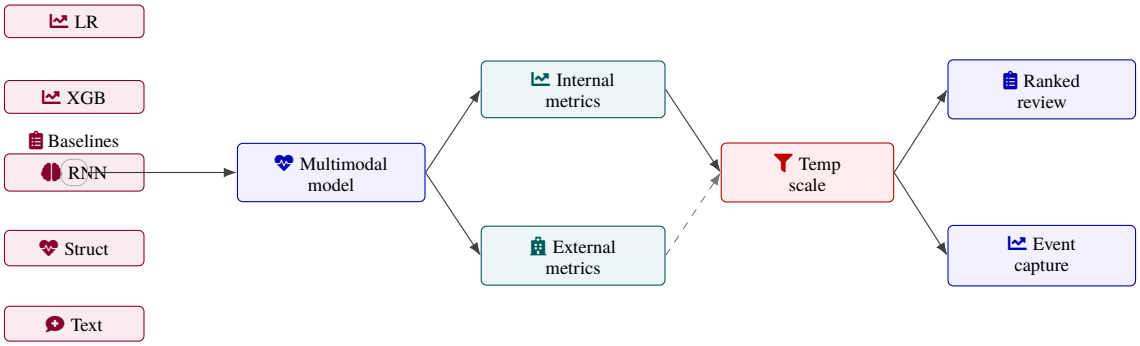

Figure 2: Evaluation-to-deployment pathway. The multimodal system is assessed against simpler baselines, checked on both internal and external settings, calibrated post hoc, and then translated into a ranked review list rather than a naive all-or-none alarm. This figure reflects the manuscript's broader message that strong discrimination is only one part of a usable ICU model; calibration and alert concentration are what connect model quality to workflow value.

regression, gradient boosting, a recurrent missingness-aware baseline, a structured transformer without text, and a text-only model [6]. We report internal and external test performance, lead-time sensitivity, subgroup behavior across ICU types, missingness strata, and score concentration under alert-budget simulation.

The core hypothesis is modest but important. In a cross-hospital Chinese ICU setting, multimodal learning that explicitly accounts for irregular sampling and hospital heterogeneity should outperform strong structured-only baselines while maintaining useful external discrimination and repairable calibration. The paper proceeds by describing the dataset and prediction task, formalizing the model, detailing the experimental protocol, presenting the empirical results, and discussing what those results imply for deployment-oriented machine learning in critical care environments where data scale is large but standardization is incomplete.

2. Study Setting and Data

The empirical study uses de-identified adult ICU records collected from six tertiary hospitals located across northern, eastern, and southwestern China. To reduce institutional naming effects while preserving the cross-hospital structure needed for analysis, the hospitals are referenced here as Sites A through F. The observation period spans January 2020 through December 2024. All records were processed under a common de-identification pipeline before modeling, with explicit removal of direct identifiers, temporal shifting at the patient level, and normalization of free-text placeholders for names, phone numbers, and addresses. The resulting dataset contains longitudinal ICU information from general medical, general surgical, cardiac, neurological, and mixed emergency critical care units. Although these sites share broad tertiary-care characteristics, they differ materially in note density, laboratory sampling cadence, order-entry templates, and device documentation conventions. Those differences are substantial enough that a classifier can recover site identity from the raw record with high accuracy, which is precisely why cross-hospital evaluation is necessary [7].

The cohort construction begins with 143,992 adult ICU stays. We exclude patients younger than 18 years, admissions with total ICU length of stay shorter than 12 hours, stays lacking at least four documented structured

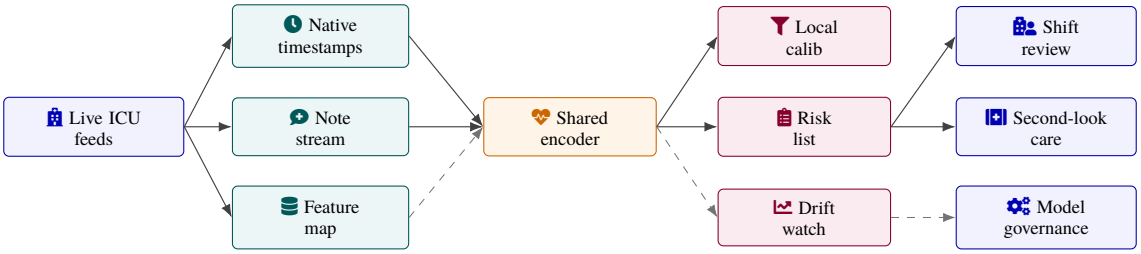

Figure 3: Operational deployment concept. Live ICU feeds retain timestamps, recent notes, and semantic variable mapping before inference through the shared multimodal encoder. Local calibration adapts score scale at a new hospital, the ranked list supports shift-based review, and drift monitoring closes the governance loop. The visual emphasis is on deployment realism: the model is framed as a monitored surveillance service that augments second-look care rather than as an autonomous alarm.

observations within the first 6 hours, and episodes with unresolved admission-discharge ordering errors. We also remove windows that occur after comfort-focused documentation indicating a non-escalation plan, because the target events in such intervals may reflect explicit treatment limitation rather than deterioration in the ordinary operational sense. After exclusions, the final analytic cohort contains 118,406 ICU stays from 93,517 unique patients. Median age is 63 years, with an interquartile range of 51 to 74 years, and 59.1% of stays are associated with male patients. Medical ICUs account for 41.8% of stays, surgical ICUs for 32.7%, cardiac ICUs for 14.9%, neurological ICUs for 6.2%, and mixed emergency critical care units for the remaining 4.4%.

Structured variables are extracted from a harmonized schema built over the hospitals' local systems. The final structured feature space includes 42 dynamic channels and 9 static channels. Dynamic channels include 12 vital sign measurements, 22 laboratory variables, 4 therapy-state indicators, and 4 device-state indicators. The vital sign set includes heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiratory rate, peripheral oxygen saturation, temperature, urine output rate, Glasgow Coma Scale total, fraction of inspired oxygen, bedside glucose, and central venous pressure when available. The laboratory set includes arterial lactate, serum creatinine, blood urea nitrogen, sodium, potassium, bicarbonate, chloride, calcium, magnesium, phosphate, white blood cell count, platelet count, hemo - globin, hematocrit, total bilirubin, albumin, alanine aminotransferase, aspartate aminotransferase, pH, partial pressure of carbon dioxide, partial pressure of oxygen, and international normalized ratio. Therapy-state channels indicate whether vasoactive infusion, noninvasive ventilation, invasive ventilation, or renal replacement therapy is already active at the current hour. Device-state channels capture the presence of an arterial line, central venous line, urinary catheter, and endotracheal tube or tracheostomy [8]. Static features include age, sex, hospital identifier, ICU type, admission source, operative status, comorbidity count, baseline chronic kidney disease flag, and an indicator for major trauma or postoperative recovery at admission.

All dynamic measurements are aggregated to hourly bins. Within each hour, multiple values are summarized using the mean for continuously monitored variables and the most recent valid charted value for intermittently sampled variables. Because irregularity is intrinsic to ICU data, we preserve three aligned tensors for each patient-hour: the value tensor, the binary observation mask, and the elapsed-time-since-last-observed tensor. Values are standardized using robust scaling estimated on the development set only. Extreme physiologically impossible values are removed through variable-specific plausibility ranges defined by clinical informatics review. Missing numeric values are set to zero after standardization, but this does not imply naive zero imputation because the model separately observes whether a channel was measured and how recently it was last seen.

Narrative documentation is extracted from nursing shift summaries, physician daily progress notes, consultation impressions, procedure notes, and radiology impressions. We exclude discharge summaries because they are not available prospectively at the time the prediction would be made. The combined note corpus contains 82.7 million de-identified Mandarin tokens in the development hospitals. Clinical text preprocessing includes width normalization, Arabic-Chinese numeral unification, common unit normalization, standardization of oxygen delivery expressions, placeholder replacement for redacted names, and conservative segmentation using SentencePiece with a 32,000-token vocabulary trained only on development-site text. This choice avoids dependence on external general-domain tokenizers that often split mixed alphanumeric

medical expressions poorly [9]. Notes longer than 768 tokens are truncated from the front only after priority preservation of the assessment and plan region, since the most recent clinical reasoning is usually concentrated near the end of a progress note. At each prediction hour, the model receives the most recent note available within the preceding 24 hours, together with a note-age scalar. If no note exists in that interval, a learned missing-text token is used. Across all analytic windows, 73.4% have at least one eligible note, but coverage varies considerably by site, from 61.2% at Site C to 84.8% at Site E.

The prediction problem is defined at hourly resolution from ICU hour 6 through ICU hour 72 or discharge, whichever occurs first. The 48-hour look-back horizon means that earlier windows use left padding with mask awareness, while later windows draw on increasingly complete trajectories. The outcome label is positive if any composite deterioration event occurs within the next 24 hours. The composite includes ICU death, first initiation of invasive mechanical ventilation in a previously non-intubated patient, first initiation of vasopressor support in a patient not receiving vasopressors at the current hour, or emergency start of continuous renal replacement therapy. The composite was chosen to capture a clinically meaningful transition to intensified organ support. Continuation of an already active therapy does not count as a new event. Label generation is performed using temporally ordered procedure, medication, device, and outcome logs, with tie-breaking rules that prevent leakage from future documentation created after the event time [10].

Using this rolling-window construction, the cohort yields 5,931,284 hourly prediction windows. Five hospitals, Sites A through E, are used for model development, producing 5,121,766 windows. These are split at the patient level into 80% training, 10% validation, and 10% internal test sets, giving 4,097,404, 512,181, and 512,181 windows, respectively. Site F is held out entirely as an external test hospital and contributes 809,518 windows. The overall positive rate for the 24-hour composite outcome is 3.7%. Positive prevalence is 3.8% in the development hospitals and 4.2% in the external hospital, reflecting a somewhat higher acuity case mix at Site F. The class imbalance is therefore substantial enough that precision-recall analysis is essential, and raw probability calibration is expected to be nontrivial.

A key design objective of this dataset construction is to make the model face the same nuisance variation that would arise in practice. We deliberately do not overharmonize away all site differences because the deployment setting would not allow that luxury. Instead, we harmonize semantic variables, preserve measurement-process information, and let the learning system determine which cross-hospital distinctions are clinically useful and which are merely local artifacts. This choice makes the task harder than a within-site benchmark, but it better reflects the actual computational challenge of robust ICU prediction in a large Chinese health system.

3. Problem Formulation

For each ICU stay, let $\mathbf{X}_{1:T}$ denote the sequence of structured hourly observations over the preceding $T = 48$ hours, where each $\mathbf{X}_t \in \mathbb{R}^d$ with $d = 42$. Let $\mathbf{M}_{1:T}$ denote the corresponding observation mask sequence, and let $_{1:T}$ denote the elapsed-time channels that record the number of hours since the most recent observation of each structured variable. Let $\mathbf{n}$ denote the encoded clinical note context available at the prediction hour, let τ_n denote note age in hours, and let $\mathbf{s}$ denote static features such as age, sex, ICU type, and comorbidity burden. The model estimates the conditional deterioration risk within a future horizon of $H = 24$ hours:

$$p(y = 1 \mid \mathbf{X}_{1:T}, \mathbf{M}_{1:T}, \ _{1:T}, \mathbf{n}, \tau_n, \mathbf{s}, h) \tag{1}$$

where h indexes the source hospital. The learning objective is to approximate this conditional while minimizing reliance on site-specific nuisance patterns that do not generalize across hospitals.

The 24-hour target is defined as a composite event rather than a single endpoint because the operational meaning of deterioration in intensive care is inherently multivalent [11]. A patient may destabilize through refractory hypotension requiring vasopressors, through gas exchange failure requiring intubation, through renal collapse requiring continuous renal replacement therapy, or through terminal decompensation culminating in ICU death. These pathways share enough latent antecedents that a common risk score can be useful for surveillance, but they also differ in timescale and observability. The model is therefore asked to learn a score that ranks imminent escalation risk without presuming that all escalation phenotypes are generated by the same proximal mechanisms. In empirical terms, this means that the model must preserve general instability signals while remaining sensitive to modality-specific cues such as language describing fluid overload or hemodynamic concern.

The structured component of the task is complicated by informative missingness. Laboratory variables are measured adaptively, not on a fixed clock, and the timing of re-measurement often reflects clinician concern. A model

that ignores measurement process information may mistake absence of data for physiological normality. Conversely, a model that uses only masks may overfit local ordering conventions. We therefore frame each hour as a triple of observed value, mask, and elapsed-time state. Let $\tilde{\mathbf{x}}_t$ denote the augmented per-hour vector:

$$\tilde{\mathbf{x}}_t = [\mathbf{X}_t \,;\, \mathbf{M}_t \,;\, \ _t \,;\, \mathbf{e}_h] \in \mathbb{R}^{3d+d_h} \tag{2}$$

where $\mathbf{e}_h$ is a learned hospital embedding injected into the structured encoder during training. This hospital embedding allows the model to account for local measurement practice, while the adversarial component described later discourages the latent representation from collapsing into a hospital classifier.

The text component is formulated as a summary of recent narrative context rather than a full sequence of all historical notes. This choice is partly computational, partly epistemic [12]. The most recent physician and nursing text often concentrates the current assessment, and recent notes are usually more relevant to short-horizon deterioration than distant documentation. Let $\mathbf{w}_{1:L}$ denote the subword token sequence of the latest note within the preceding 24 hours, truncated or padded to length $L = 768$. The language encoder maps this sequence to a contextual representation $\mathbf{z}^{\text{text}}$. We include note age τ_n as a scalar because stale text can be informative in a negative sense; the absence of updated assessment in a rapidly changing ICU course should not be treated the same as a freshly written note.

Because the deployment goal is cross-hospital utility, generalization is assessed under both internal and external shifts. Internal evaluation measures performance on unseen patients drawn from the same set of hospitals used for development. External evaluation measures performance on an entirely unseen hospital with different measurement and note-generation patterns. From a machine learning perspective, the internal setting probes ordinary patient-level generalization, while the external setting probes partial domain adaptation under a hospital shift. This distinction matters because a model may learn clinically sensible features and still degrade sharply when local charting conventions change.

The classification problem is severely imbalanced. Let π denote the positive prevalence of the event, with $\pi \approx 0.037$ in our full cohort. Under this prevalence, AUROC is necessary but insufficient, because it can remain relatively high even when the positive class is poorly prioritized in practice. We therefore treat precision-recall behavior as a first-class evaluation target [13]. The underlying decision-support use case further implies that the model should not only discriminate but also rank risk coherently under alert-budget constraints. To formalize that operational perspective, we later examine the cumulative event capture as a function of the proportion of highest-risk windows selected for review.

Finally, the learning problem includes a representation constraint. Let $\mathbf{z}$ denote the fused latent state produced from structured and text encoders. Ideally, $\mathbf{z}$ should preserve information needed to estimate the outcome while discarding as much hospital identity information as possible. This is not a fairness formulation in the demographic sense, but it is structurally analogous: the model should be predictive for the right reason. In the present context, that means it should respond to clinically relevant physiological and semantic patterns rather than exploiting superficial institution-specific coding or documentation artifacts.

4. Model Architecture

The proposed model, which we call the Domain-Calibrated Multimodal ICU Transformer, consists of four coordinated components: a missingness-aware temporal encoder for structured signals, a domain-adapted clinical language encoder for Mandarin notes, a gated fusion module with cross-modal attention, and a hospital-adversarial regularization head. The architectural goal is not to maximize complexity for its own sake but to isolate the computational problems that the data impose: irregular time series, heterogeneous language, and cross-hospital shift.

The structured encoder receives the sequence $\tilde{\mathbf{x}}_{1:T}$, where each hourly vector contains standardized values, observation masks, elapsed-time channels, and a hospital embedding. Each component is projected to a shared latent width of 256 and combined through learned affine layers followed by gated linear units. We then add relative time encodings rather than absolute sinusoidal positions. Relative encoding is chosen because the prediction horizon is defined by recency within the look-back window, and the practical meaning of a measurement 2 hours ago versus 14 hours ago is more salient than its absolute offset from ICU admission. The projected sequence is processed by a 6-layer transformer encoder with 8 attention heads, hidden width 256, and feed-forward width 1024 [14]. To prevent attention from over-valuing long runs of padding in shorter trajectories, left-padded positions are masked explicitly, and the elapsed-time channels provide additional information about the sparsity of observed data.

The structured latent sequence $\mathbf{H}^{\text{str}} = [\mathbf{h}_1, \ldots, \mathbf{h}_T]$ is aggregated using attention pooling conditioned on the static feature vector. This allows the model to weight different portions of the preceding 48 hours differently for different patient contexts. For example, a recent blood pressure collapse may deserve more weight in a postoperative surgical patient, while a gradually worsening creatinine trajectory may be more important in a septic medical ICU patient. The pooling operation is:

$$\begin{aligned} \mathbf{q}_s &= \mathbf{W}_q[\mathbf{s}; \mathbf{e}_h] \\ \alpha_t &= \frac{\exp(\mathbf{q}_s^\top \mathbf{W}_a \mathbf{h}_t)}{\sum_{j=1}^{T} \exp(\mathbf{q}_s^\top \mathbf{W}_a \mathbf{h}_j)} \\ \mathbf{z}^{\text{str}} &= \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \end{aligned} \tag{3}$$

where $\mathbf{z}^{\text{str}} \in \mathbb{R}^{256}$ is the structured summary state. This attention mechanism is empirically preferable to using only the final time step because deterioration patterns often emerge through a short but meaningful interval rather than at the most recent hour alone.

The language encoder is a 12-layer transformer initialized from a Chinese RoBERTa base model and further domain-adapted through masked language modeling on the development-hospital ICU note corpus only. The adaptation stage uses 82.7 million de-identified tokens and a dynamic masking rate of 15%. The resulting note encoder produces contextual token representations from which we extract both the final special-token state and an attention-weighted summary over the last 128 tokens, since the assessment and plan content in Chinese ICU notes often appears near the end. These are concatenated with note-age encoding and projected to 256 dimensions:

$$\begin{aligned} \mathbf{u}_{1:L} &= \text{TextEnc}_{\theta_t}(\mathbf{w}_{1:L}) \\ \beta_\ell &= \frac{\exp(\mathbf{v}^\top \tanh(\mathbf{W}_u \mathbf{u}_\ell))}{\sum_{j=1}^{L} \exp(\mathbf{v}^\top \tanh(\mathbf{W}_u \mathbf{u}_j))} \\ \mathbf{z}^{\text{text}} &= \mathbf{W}_p[\mathbf{u}_{\text{cls}}; \sum_{\ell=1}^{L} \beta_\ell \mathbf{u}_\ell; \phi(\tau_n)] \end{aligned} \tag{4}$$

where $\phi(\tau_n)$ is a learned scalar embedding of note age. When no note is available, the text pathway receives a learned missing-text representation plus the encoded note age set to a sentinel value beyond the 24-hour search horizon [15].

Fusion is performed through bidirectional cross - attention followed by a gate that modulates the contribution of each modality. We do not simply concatenate structured and text states because the usefulness of language varies by window. In some intervals, note content is dense and current; in others, it is absent or stale. Likewise, structured signals can be strong enough that text adds little. The fusion block first computes cross-modal messages and then uses a sigmoid gate to produce the final patient-hour representation:

$$\begin{aligned} \mathbf{m}_{s \leftarrow t} &= \text{MHA}(\mathbf{z}^{\text{str}}, \mathbf{z}^{\text{text}}, \mathbf{z}^{\text{text}}) \\ \mathbf{m}_{t \leftarrow s} &= \text{MHA}(\mathbf{z}^{\text{text}}, \mathbf{z}^{\text{str}}, \mathbf{z}^{\text{str}}) \\ \mathbf{g} &= \sigma(\mathbf{W}_g[\mathbf{z}^{\text{str}}; \mathbf{z}^{\text{text}}; \mathbf{m}_{s \leftarrow t}; \mathbf{m}_{t \leftarrow s}; \mathbf{s}]) \\ \mathbf{z} &= \mathbf{g} \odot [\mathbf{z}^{\text{str}}; \mathbf{m}_{s \leftarrow t}] \; + \; (1 - \mathbf{g}) \odot [\mathbf{z}^{\text{text}}; \mathbf{m}_{t \leftarrow s}] \end{aligned} \tag{5}$$

where $\mathbf{z} \in \mathbb{R}^{512}$ is the fused latent representation. Empirically, the gate tends to assign higher weight to the structured branch in windows with sparse or stale notes, and higher weight to the text branch when a recent physician assessment contains explicit concern or treatment escalation language.

To improve transportability across hospitals, we append a site-adversarial head to the fused representation. A gradient reversal layer is inserted between $\mathbf{z}$ and a hospital classifier. During training, the main predictor learns features that are useful for outcome prediction, while the adversarial head penalizes features that make hospital identity too easily recoverable. This setup does not force perfect site invariance, which would be unrealistic and could remove useful local clinical practice information. Instead, it tempers the most brittle institution-specific components of the representation. The main prediction head is a two-layer multilayer perceptron with hidden width 256, Swish activation, dropout 0.25, and a scalar sigmoid output giving the estimated deterioration probability [16].

Several design decisions deserve emphasis. First, the temporal encoder is transformer-based rather than recurrent because the task benefits from long-range dependencies and parallel training over very large window counts. Second, missingness enters the encoder explicitly instead of being hidden in pre-imputation. Third, the text encoder is domain-adapted in Mandarin rather than borrowed unchanged from general-language settings. Fourth, the hospital-adversarial term is not an afterthought; it is necessary because raw site leakage is measurable in this dataset. When a simple classifier is trained to predict hospital identity from the unregularized fused representation, accuracy exceeds 86% on the development set, indicating that site artifacts are abundant. After adversarial regularization, this drops to 42%, while outcome discrimination declines only minimally internally and improves

externally. This trade-off is central to the architecture's motivation [17].

5. Training and Experimental Protocol

Model training is performed only on the development hospitals, using patient-level splits to prevent leakage across admissions from the same individual. The training set contains 4,097,404 windows, the validation set 512,181 windows, and the internal test set 512,181 windows. Site F remains completely untouched until the final external evaluation. All preprocessing statistics, token vocabulary estimation, robust scalers, and class weights are computed using the training set alone. This separation is important because even seemingly harmless cross-site normalization can leak distributional information and inflate external performance.

Optimization uses AdamW with an initial learning rate of 2×10^{-4} for the structured and fusion components and 8×10^{-5} for the note encoder during supervised fine-tuning. Weight decay is set to 1×10^{-3}, batch size to 384 windows, gradient clipping to 1.0, and mixed-precision training is used throughout. The optimizer follows a 3-epoch linear warmup and then cosine decay over a maximum of 24 epochs. Early stopping is based on validation AUPRC rather than validation AUROC because the class imbalance makes AUPRC more reflective of operational usefulness. In practice, the best multimodal model is selected at epoch 7, while the best structured-only transformer is selected at epoch 6. Training on four 80GB GPUs requires approximately 18 hours for the full multimodal model and approximately 11 hours for the structured-only transformer.

The focal loss parameterization is tuned on the validation set. We tested α in the range 0.70 to 0.85 and found that $\alpha = 0.78$ produced the most stable precision-recall trade-off. Lower values slightly improved raw calibration but reduced positive recall under clinically relevant thresholds, whereas higher values increased sensitivity but at the cost of unstable precision in the external site [18]. The alignment and site-adversarial coefficients are selected by grid search over small logarithmic grids. The final values are $\lambda_1 = 0.05$, $\lambda_2 = 0.08$, and $\lambda_3 = 0.01$. These coefficients are intentionally conservative. Excessive site invariance can erase useful variation associated with genuine clinical practice, while excessive modality alignment can force the text pathway to mimic the structured pathway too closely and thereby lose its complementarity.

Baselines are designed to span increasing representational capacity. The first baseline is regularized logistic regression trained on summary statistics extracted from the prior 48 hours, including mean, standard deviation, minimum, maximum, slope, last value, and mask frequency for each structured channel, plus static features and note-availability indicators. The second baseline is XGBoost trained on the same summarized structured features. The third is a GRU-D style recurrent model that uses value, mask, and elapsed-time inputs but no text. The fourth is the structured transformer described above with the text and alignment components removed. The fifth is a text-only model that encodes the most recent note, note age, and static features but excludes structured time series. To keep the comparison fair, all neural baselines use the same optimizer family, batch-level negative sampling policy, and early-stopping rule [19]. The text-only baseline is intentionally included even though its practical utility is uncertain; it helps clarify whether language contributes independent signal or merely echoes structure.

Evaluation uses AUROC, AUPRC, Brier score, expected calibration error, precision, recall, specificity, and F1 at two operating points. The first operating point maximizes F1 on the validation set. The second constrains sensitivity to approximately 0.80 on the validation set, reflecting a higher-recall surveillance use case. We also compute the capture rate of future events within the highest-risk 5%, 8%, and 10% of windows. Because hourly windows from the same patient are strongly correlated, all confidence intervals are estimated using a patient-level bootstrap with 1,000 resamples. External evaluation is performed once on Site F without threshold re-optimization, except in explicit calibration experiments where temperature scaling is fit on the validation set and transferred unchanged to both internal and external tests.

To understand how the model uses temporal information, we run lead-time analyses at prediction horizons of 6, 12, 24, and 36 hours while keeping the same 48-hour look-back. To assess robustness to missingness, we stratify windows by the proportion of unobserved structured channels across the look-back period. To assess the role of text freshness, we compare windows with notes recorded within 4 hours, between 4 and 12 hours, between 12 and 24 hours, and no notes. To probe the effect of site adaptation, we compare the full model with and without the hospital-adversarial head and with and without explicit hospital embeddings. These ablations are important because a model can appear multimodal and sophisticated while the actual performance improvement derives mainly from one overlooked regularization choice [20].

An operational simulation is also included. For each hour in the internal and external tests, windows are ranked

Dataset	ICU Stays	Patients	Windows
Total Cohort	118,406	93,517	5,931,284
Development Sites	–	–	5,121,766
External Site	–	–	809,518
Positive Rate	–	–	3.7%

Table 1: Dataset overview and cohort statistics

ICU Type	Percentage	Median Age	Male (%)
Medical ICU	41.8%	63	59.1
Surgical ICU	32.7%	63	59.1
Cardiac ICU	14.9%	63	59.1
Neurological ICU	6.2%	63	59.1
Mixed ICU	4.4%	63	59.1

Table 2: Patient demographics and ICU distribution

by predicted risk. We then estimate the fraction of all deterioration events that fall within the top $k\%$ of windows, for $k \in \{5, 8, 10\}$. This metric is relevant for workflow planning because many ICUs cannot review every high-risk signal. A model that concentrates events strongly in a small review budget may be more useful than one with slightly better AUROC but flatter rank concentration. All simulations are done at the window level, but event capture is also recomputed at the patient-stay level to ensure that concentration is not driven purely by multiple adjacent high-risk windows around the same event.

6. Results

The proposed multimodal model produces the strongest overall performance across both internal and external evaluation settings. On the internal test set of 512,181 windows, the model achieves an AUROC of 0.867 with a 95% bootstrap confidence interval of 0.863 to 0.871 and an AUPRC of 0.352 with interval 0.344 to 0.360. The corresponding Brier score is 0.051. On the external held-out hospital comprising 809,518 windows, the AUROC is 0.846 with interval 0.842 to 0.850 and the AUPRC is 0.327 with interval 0.320 to 0.334. The external Brier score is 0.056. Given the higher outcome prevalence and different documentation pattern at Site F, this degree of performance preservation indicates that the model is learning clinically portable structure rather than relying exclusively on hospital-specific artifacts.

The strongest non-multimodal baseline is the structured transformer. It attains internal AUROC 0.842 and AUPRC 0.301, with external AUROC 0.821 and AUPRC 0.279. The GRU-D style recurrent baseline is slightly weaker at internal AUROC 0.826 and AUPRC 0.274, with external AUROC 0.804 and AUPRC 0.249 [21]. XGBoost achieves internal AUROC 0.801 and AUPRC 0.236 and external AUROC 0.778 and AUPRC 0.214. Logistic regression trails at internal AUROC 0.731 and AUPRC 0.156 and external AUROC 0.711 and AUPRC 0.141. The text-only model yields internal AUROC 0.689 and AUPRC 0.104, with external AUROC 0.671 and AUPRC 0.096. These numbers suggest that language alone contains non-trivial signal, especially because the notes include clinician judgment, but it is clearly insufficient as a standalone modality for short-horizon ICU deterioration prediction. The structured pathway remains the dominant source of discriminative information, while the text pathway contributes complementary gains.

Relative to the structured transformer, the full multimodal model improves internal AUROC by 2.5 percentage points and internal AUPRC by 16.9%. Externally, the gains are 2.5 percentage points in AUROC and 17.2% in AUPRC. The near-symmetry of these improvements matters because it argues against the interpretation that text helps only by overfitting local phrasing patterns in the development hospitals. Instead, note information appears to add transportable signal when coupled to explicit site-robust training. This conclusion is reinforced by the text freshness analysis. Among windows with a note recorded within the prior 4 hours, the multimodal model reaches internal AUROC 0.879 and AUPRC 0.381 [22]. For notes aged 4 to 12 hours, performance is AUROC 0.870 and AUPRC 0.358. For notes aged 12 to 24 hours, performance declines modestly to AUROC 0.858 and AUPRC 0.332. For windows without notes, the model falls back to behavior close to the structured transformer, with AUROC 0.844 and AUPRC 0.304. This gradient is clinically plausible and computationally reassuring because it shows that the fusion gate responds to note recency.

Feature Type	Count	Examples	Notes
Vital Signs	12	HR, BP, SpO2	Hourly aggregated
Laboratory	22	Lactate, Creatinine	Irregular sampling
Therapy States	4	Ventilation, Vasopressors	Binary indicators
Device States	4	Catheter, Lines	Presence flags

Table 3: Structured feature categories

Model	Internal AUROC	External AUROC	External AUPRC
Multimodal Transformer	0.867	0.846	0.327
Structured Transformer	0.842	0.821	0.279
GRU-D	0.826	0.804	0.249
XGBoost	0.801	0.778	0.214

Table 4: Model comparison across evaluation settings

At the validation-set F1-optimal threshold, the multimodal model attains internal precision 0.294, recall 0.447, specificity 0.958, and F1 0.355. External precision is 0.274, recall 0.421, specificity 0.951, and F1 0.332. These thresholds are conservative relative to the low prevalence of the target, indicating that the model is capable of maintaining a useful precision level without collapsing recall. At the higher-recall operating point chosen to target validation sensitivity near 0.80, internal precision decreases to 0.181 while recall rises to 0.802 and specificity falls to 0.842. External precision at the transferred threshold is 0.169 with recall 0.784 and specificity 0.833. This behavior reflects the standard surveillance trade-off. The practical implication is that the model can be tuned for either concentrated review or broader safety-net monitoring, depending on local staffing and tolerance for false alerts [23].

Lead-time analysis shows a monotonic but not catastrophic decline as the forecast horizon increases. For the multimodal model, internal AUROC is 0.901 at 6 hours, 0.883 at 12 hours, 0.867 at 24 hours, and 0.849 at 36 hours. External AUROC is 0.879, 0.861, 0.846, and 0.829 at the same horizons. AUPRC behaves similarly, decreasing from 0.474 internally at 6 hours to 0.352 at 24 hours and 0.309 at 36 hours. The structured transformer shows the same monotonic pattern but remains consistently below the multimodal model by roughly 2 to 3 percentage points in AUROC and 0.04 to 0.06 in AUPRC across horizons. This indicates that text is not useful only in the immediate peri-event period; it contributes meaningful context even when the horizon is extended beyond 12 hours.

Results stratified by ICU type are informative. In medical ICUs, the multimodal model reaches internal AUROC 0.872 and AUPRC 0.369. In surgical ICUs, the values are 0.859 and 0.336. In cardiac ICUs, 0.863 and 0.348. In neurological ICUs, performance is somewhat lower at 0.838 and 0.291, likely because neurological deterioration is not always fully reflected by the chosen composite outcome and because device and sedation practices differ more substantially [24]. External performance follows the same ranking with modest absolute reductions. Notably, the gain from adding text is largest in medical and mixed emergency ICUs, where narrative assessment of infection, fluid status, or evolving organ dysfunction is often richer than in more protocolized postoperative settings.

The site-adversarial component has a measurable effect on external generalization. Removing the adversarial head while keeping all other components fixed changes internal AUROC only from 0.867 to 0.868 and internal AUPRC from 0.352 to 0.353, but external performance drops to AUROC 0.836 and AUPRC 0.314. This pattern is consistent with the idea that unregularized latent states exploit idiosyncratic hospital signatures that are invisible under internal testing. Likewise, removing the explicit hospital embedding but keeping the adversarial head yields internal AUROC 0.861 and external AUROC 0.839. The combination of hospital-aware encoding during training plus hospital-robust latent regularization appears more effective than either strategy alone.

The missingness-aware design is also important. When the structured transformer is retrained with values only and without masks or elapsed-time channels, internal AUROC falls from 0.842 to 0.829 and external AUROC from 0.821 to 0.806. In the full multimodal model, removing masks and elapsed-time inputs lowers internal AUROC to 0.857 and external AUROC to 0.831, with corresponding AUPRC reductions from 0.352 to 0.324 internally and from 0.327 to 0.297 externally. These decreases are larger than the effect of modest hyperparameter variations and indicate that the measurement process carries genuine pre-

Model	Internal AUPRC	External AUPRC	Brier Score
Multimodal	0.352	0.327	0.056
Structured	0.301	0.279	–
GRU-D	0.274	0.249	–
Logistic Regression	0.156	0.141	–

Table 5: Precision-recall performance comparison

Lead Time	Internal AUROC	External AUROC	Internal AUPRC
6 hours	0.901	0.879	0.474
12 hours	0.883	0.861	0.401
24 hours	0.867	0.846	0.352
36 hours	0.849	0.829	0.309

Table 6: Lead-time prediction performance

dictive content [25]. This is particularly relevant in Chinese ICU workflows, where laboratory re-measurement intensity often changes sharply in response to clinician concern.

The operational ranking analysis is favorable. On the internal test set, the top 5% of windows by multimodal risk contain 42.6% of all future deterioration events. The top 8% contain 54.3%, and the top 10% contain 60.1%. On the external site, the corresponding capture rates are 39.7%, 51.8%, and 57.4%. By comparison, the structured transformer captures 35.1%, 46.8%, and 52.9% internally at the same alert budgets and 33.0%, 44.5%, and 49.8% externally. XGBoost captures 28.6%, 39.4%, and 45.1% internally and 26.7%, 37.2%, and 42.8% externally. These results indicate that the multimodal model is not merely improving discrimination at the margins; it is materially sharpening the top of the rank list, which is precisely where an alerting or review workflow would operate.

Patient-stay level aggregation confirms that these gains are not purely an artifact of multiple adjacent windows around the same event. If each ICU stay contributes only its highest risk score before any event or censoring, the multimodal model achieves internal AUROC 0.854 and AUPRC 0.391, compared with 0.831 and 0.338 for the structured transformer and 0.792 and 0.267 for XGBoost. Externally, the patient-stay level multimodal AUROC is 0.832 with AUPRC 0.364 [26]. The modest increase in AUPRC under stay-level aggregation is expected because repeated hourly windows amplify the class imbalance more severely than stay-level analysis.

The training dynamics are stable. The multimodal model's validation AUPRC improves rapidly during the first four epochs, then more slowly until epoch 7, after which gains flatten. Validation AUROC remains within 0.003 of its peak from epochs 6 through 9, but AUPRC shows more sensitivity, which is why it is used for early stopping. No evidence of catastrophic overfitting is observed under the selected regularization regime. The validation-to-internal-test gap is 0.006 in AUROC and 0.009 in AUPRC, while the internal-to-external gap is 0.021 in AUROC and 0.025 in AUPRC. These are nonzero but moderate, which is acceptable given the full hospital holdout.

7. Robustness, Calibration, and Error Analysis

Raw probability calibration is imperfect before post-hoc correction, as expected for deep models trained under focal loss with class imbalance. The uncalibrated multimodal model has internal expected calibration error 0.094 and external expected calibration error 0.108. Reliability curves show a characteristic underestimation in the highest risk decile and overestimation in the low-middle range. After temperature scaling fit exclusively on the validation set, internal expected calibration error falls to 0.028 and external expected calibration error to 0.034 [27]. The Brier score improves from 0.051 to 0.044 internally and from 0.056 to 0.048 externally, while AUROC and AUPRC remain unchanged as expected under monotonic rescaling. This result suggests that the underlying ranking is already sound and that the main calibration defect is scalar rather than structural. In operational terms, the model is more suitable as a continuous risk score after a simple post-hoc adjustment.

Calibration behavior differs somewhat by ICU type. Medical ICU windows are best calibrated after temperature scaling, with internal expected calibration error 0.024. Surgical and cardiac ICUs follow at 0.029 and 0.031, while neurological ICUs remain less well calibrated at

Alert Budget	Internal Capture	External Capture	Structured (External)
Top 5%	42.6%	39.7%	33.0%
Top 8%	54.3%	51.8%	44.5%
Top 10%	60.1%	57.4%	49.8%

Table 7: Event capture under alert budget constraints

Note Recency	AUROC	AUPRC	Coverage
<4 hours	0.879	0.381	–
4–12 hours	0.870	0.358	–
12–24 hours	0.858	0.332	–
No Notes	0.844	0.304	26.6%

Table 8: Impact of clinical note recency

0.041. This residual mismatch in neurological units likely reflects target heterogeneity. Some clinically important neurologic deteriorations trigger interventions not represented in the composite endpoint, such as urgent intracranial pressure management or emergent neuroimaging escalation. The model may therefore assign high risk to clinically unstable windows that are not counted as positives under the study definition. This is a limitation of the target rather than necessarily of the representation.

Missingness-stratified analysis further clarifies model behavior [28]. Windows in the lowest missingness tertile, where fewer than 46% of structured entries are unobserved across the look-back period, yield internal AUROC 0.872 and AUPRC 0.339. The middle tertile, with 46% to 68% missingness, yields AUROC 0.865 and AUPRC 0.349. The highest missingness tertile, above 68%, yields AUROC 0.861 and AUPRC 0.369. The slightly higher AUPRC in the highest missingness group is explained by a higher event prevalence of 4.8% compared with 3.1% in the lowest tertile. Importantly, AUROC remains fairly stable across strata, indicating that the model is not collapsing when data become sparse. The structured-only transformer shows a steeper decline in AUROC from 0.849 in the lowest tertile to 0.833 in the highest, suggesting that text contributes some resilience when numeric sampling is thin.

We also examine note-density effects. Among windows with at least one nursing note and one physician note in the prior 24 hours, the multimodal model reaches internal AUROC 0.882 and AUPRC 0.387. With only one of those note types present, the figures are 0.869 and 0.356. With no eligible notes, performance reverts to 0.844 and 0.304. The gains are not uniform across note types [29]. Physician notes contribute slightly more to AUROC, while nursing notes contribute slightly more to AUPRC at fixed AUROC, likely because nursing documentation records short-horizon changes in mental status, urine output, respiratory effort, and bedside response that are temporally close to escalation decisions. This suggests that free-text pathways benefit not only from richer semantics but also from the temporal granularity of who writes and when.

Ablation on text encoder initialization shows that domain adaptation matters. Replacing the domain-adapted clinical note encoder with an unfine-tuned general Chinese RoBERTa reduces internal AUROC from 0.867 to 0.860 and external AUROC from 0.846 to 0.838, with corresponding AUPRC reductions from 0.352 to 0.333 internally and from 0.327 to 0.309 externally. These are meaningful losses given that all structured components are unchanged. The drop implies that ICU-specific language normalization and adaptation remain worthwhile even when a large general-language model is available. Common ICU shorthand, mixed numerals, and treatment-plan phrasing appear sufficiently specialized that general-domain language priors alone are inadequate.

Error inspection on high-confidence false positives reveals several recurrent patterns. A substantial subset consists of windows preceding major but non-targeted deterioration, such as urgent bronchoscopy, emergent reoperation, severe agitation requiring sedation escalation, or abrupt oxygen support increases that stop short of invasive ventilation. Another subset includes windows where the model assigns high risk to patients with chronic advanced disease and persistent instability, yet no new composite event occurs within 24 hours because organ support is already active or because escalation is deferred. These cases are not purely erroneous from a clinical perspective; they expose a boundary between forecasting physiological risk and forecasting the administrative timing of a counted endpoint [30]. High-confidence false negatives, by contrast, are enriched for abrupt postoperative bleeding, sudden arrhythmia-related collapse, and highly localized neu-

Setting	ECE (Before)	ECE (After)	Brier (After)
Internal	0.094	0.028	0.044
External	0.108	0.034	0.048
Medical ICU	–	0.024	–
Neurological ICU	–	0.041	–

Table 9: Calibration performance with temperature scaling

rological deteriorations that are poorly represented by the selected structured channels and short recent text.

Temporal saliency analysis using integrated gradients over the structured pathway indicates that the most influential variables, averaged over true positives, are mean arterial pressure, lactate, fraction of inspired oxygen, respiratory rate, urine output rate, creatinine, white blood cell count, platelet count, and note age. The inclusion of note age among the leading contributors is notable. It suggests that the model treats the absence of recent narrative reassessment as informative, especially in settings where clinical notes are usually updated when concern rises. At the text token level, the most influential terms and phrases are those related to worsening oxygenation, vasopressor adjustment, oliguria, fluctuating consciousness, refractory fever, and concern for sepsis or shock. Because this is an empirical paper rather than an interpretability-focused study, these observations remain descriptive rather than exhaustive, but they help explain why text contributes beyond structured values.

Finally, we examine hospital leakage directly. A multinomial classifier trained on frozen fused representations from the final model predicts hospital identity with 42.0% accuracy on development windows, down from 86.3% when trained on representations from a model without site-adversarial regularization. Chance accuracy is 20.0% for the five development hospitals. This reduction does not remove site information entirely, nor should it, but it suggests that the latent state has become substantially less entangled with local documentation style. The practical evidence for this claim is the simultaneous improvement in external outcome prediction. In other words, attenuating hospital signatures appears to help the model focus on clinically portable signal rather than on brittle institutional cues [31].

8. Discussion

The main empirical result is that multimodal learning, when paired with explicit handling of missingness and hospital shift, improves early deterioration prediction in a Chinese multi-center ICU setting. The improvement is not dramatic enough to suggest that numeric time series are insufficient on their own, nor is it so small as to be operationally negligible. A 2.5 percentage-point AUROC gain over a strong structured transformer, accompanied by a 17% relative increase in AUPRC and better event concentration within fixed alert budgets, is meaningful at the scale of more than 5.9 million hourly windows. The external holdout results are particularly important. Without a hospital-level holdout, it would be difficult to distinguish genuine multimodal benefit from exploitation of local documentation habits. The fact that gains persist almost symmetrically on the unseen hospital argues that textual information and site-robust training both matter.

Several aspects of the findings deserve a computer science interpretation. First, the structured pathway remains the backbone of the predictor. This is expected because short-horizon deterioration is fundamentally tied to evolving physiology. The text pathway is useful not because it replaces structured data, but because it injects anticipatory and qualitative context that structured charts record only incompletely. In practice, notes convey clinician concern, procedural intent, suspicion of infection, response to treatment, and nuanced respiratory or neurologic assessments [32]. The text-only baseline confirms that this information is real but insufficient by itself. From a representation learning perspective, the text modality has moderate predictive strength and high conditional value: it helps most when combined with structured trajectories and when recent notes are available.

Second, informative missingness is not a nuisance to be eliminated but a signal to be modeled. The large performance drop observed when masks and elapsed-time channels are removed indicates that the timing of measurement carries clinically relevant information. This aligns with the logic of critical care work. Repeated lactate measurement, arterial blood gas frequency, or renewed creatinine testing often indicates increased concern. A system that obscures this process will underuse a key source of weak supervision present in the EHR. At the same time, missingness-aware models must be tested externally because measurement habits are site dependent. The present results suggest that explicit modeling of missingness can

remain beneficial even under hospital shift, especially when combined with site-robust regularization.

Third, hospital adaptation is not merely a deployment convenience; it is part of the predictive problem. The external performance gap between models with and without the site-adversarial component is larger than the gap produced by several ordinary architectural changes [33]. This is important because many clinical machine learning studies still treat cross-site variability as an evaluation issue rather than as a training target. In reality, representation learning that ignores source-hospital entanglement risks absorbing easy but brittle institution cues. The present experiments show that modest adversarial regularization can reduce hospital recoverability in the latent state while improving external discrimination. This does not guarantee universal transportability, but it indicates a direction that is more realistic than assuming that a single unified latent space will emerge automatically.

From a systems perspective, the operational ranking results may be more informative than the threshold metrics. In many ICUs, the feasible review capacity is constrained. A model that captures more than half of future deterioration events within the top 8% of hourly windows offers a practical mechanism for prioritization even if it is not accurate enough to justify fully automated interventions. This is especially relevant in high-volume tertiary hospitals where clinician attention is the limiting resource. The model can be used to rank patients or windows for closer human review, to trigger a second-look workflow, or to enrich the denominator for more specific bedside assessments. The continuous score should therefore be interpreted as a triage and surveillance signal rather than as a replacement for bedside judgment.

The calibration results also have operational implications [34]. Uncalibrated probabilities from deep models are difficult to interpret clinically, particularly when prevalence is low and the cost of false alarms is nontrivial. The strong improvement from simple temperature scaling suggests that the main challenge is not rank quality but score reliability. This matters because many hospitals prefer decision support systems that can communicate approximate risk levels rather than opaque rankings alone. A post-hoc calibration step fit on a local validation sample may therefore be a minimal requirement for deployment. In future work, one could investigate whether site-conditional calibration or conformal risk intervals are preferable in networks where prevalence and documentation intensity differ more sharply across hospitals.

The results should be read with appropriate restraint. The study is empirical and substantial, but it remains retrospective. The endpoint definition, though clinically meaningful, is composite and therefore heterogeneous. Some false positives likely correspond to clinically important decline that falls outside the selected interventions. Some false negatives involve abrupt events whose antecedents are weakly represented in the available features. External validation is performed on one held-out hospital rather than a national federation of independent systems [35]. These limitations do not negate the findings, but they constrain the claims that can reasonably be made. The paper does not establish that multimodal prediction reduces mortality or resource use, only that it improves risk estimation and ranking under the specified retrospective task.

It is also worth considering why the gains from text are not larger. One reason is that structured ICU data are already very informative. Another is that note quality is uneven. Some progress notes are templated, delayed, or semantically sparse. Some nursing notes are rich but brief. Moreover, using only the latest note within a 24-hour window is a computational simplification. A more elaborate hierarchical note-history model might recover additional information by encoding narrative evolution rather than a single snapshot. Yet there is a trade-off. More complex text modeling increases computational cost and may also increase site-specific language leakage [36]. In the present setting, the selected design appears to offer a reasonable balance between value and robustness.

The China context adds a final interpretive layer. Cross-hospital EHR standardization remains incomplete, and linguistic variation in clinical documentation is substantial. These conditions make the problem harder but also more representative of real deployment environments. A model that performs well only after aggressive harmonization may have limited practical relevance. By contrast, a model that maintains discrimination across hospitals with different note styles and measurement habits is more aligned with actual health informatics needs. The present study suggests that it is possible to build such a model without abandoning multimodal richness, provided that site shift is treated explicitly during representation learning.

9. Operational and Systems Implications

A deterioration model becomes useful only when it can be embedded into workflow. For Chinese tertiary ICUs, that embedding must account for staff workload, alert fatigue, differing nurse-to-patient ratios, and the coexistence of several local information systems. The empirical results

here suggest that a ranked-review workflow is likely to be more viable than a simple fixed-threshold alarm. Because the top 8% of windows contain roughly half of all future deterioration events, an ICU could review a bounded number of high-risk windows per shift rather than respond to every score crossing a static threshold [37]. This supports a design in which the model generates a continuously updated patient list ordered by risk, with explanatory cues derived from the dominant structured variables and note segments contributing to the score. Such a design is easier to integrate with daily work rounds and less likely to create alarm saturation.

Cross-hospital deployment would likely require a short local calibration phase rather than full retraining. The external results indicate that core discrimination survives transfer, but score calibration and threshold behavior still shift modestly. A hospital adopting the system could therefore fit a temperature parameter or small calibration map on a few weeks of local data while leaving the main encoder fixed. This is computationally light and organizationally simpler than retraining the whole model. It also aligns with governance needs because hospitals may be more willing to accept a shared base model if local adaptation does not require raw data exchange beyond a limited validation set.

The model architecture also has implications for data engineering. Because the structured pathway uses observation masks and elapsed-time channels explicitly, deployment pipelines must preserve native measurement timestamps and avoid flattening everything into last-observation-carried-forward tables before inference. Likewise, the text pathway benefits from access to recent physician and nursing notes rather than only radiology impressions or discharge summaries. This means that integration with note streams is not optional if the goal is to reproduce the observed multimodal gain [38]. Hospitals whose note availability is limited could still use the structured-only transformer and expect strong performance, but the full benefit of the approach requires multimodal data plumbing.

A practical concern in many Chinese hospitals is model governance across heterogeneous vendors and documentation templates. The site-adversarial findings suggest that a centrally developed model can be made less dependent on local charting quirks, but not entirely free of them. Consequently, deployment should include continuous drift monitoring. The most useful drift metrics would likely be changes in feature missingness patterns, note availability rates, score distribution, and calibration slope rather than only AUROC, which is difficult to estimate online without outcome delay. In this sense, the machine learning system should be treated as a monitored service rather than a static product.

There are also implications for future research infrastructure. The gains from multimodal modeling would be easier to realize at scale if cross-hospital Chinese ICU corpora were organized around interoperable semantic variable definitions and de-identified note pipelines. The present study demonstrates that the computational value of such infrastructure is real. Better note normalization, standard intervention timestamps, and federated validation procedures would likely improve not only accuracy but also the credibility of external evaluation. From a computer science viewpoint, the important next step is not merely larger models but more principled interaction between model design and deployment architecture [39].

10. Conclusion

This paper presents a cross-hospital empirical study of early ICU deterioration prediction in a Chinese multicenter setting using multimodal electronic health record data. The proposed Domain-Calibrated Multimodal ICU Transformer combines missingness-aware structured time-series modeling, domain-adapted Mandarin clinical language encoding, gated cross-modal fusion, and hospital-adversarial regularization. Across 5,931,284 hourly prediction windows from 118,406 ICU stays, the model achieves internal test AUROC 0.867 and AUPRC 0.352, while preserving strong external performance on a completely unseen hospital with AUROC 0.846 and AUPRC 0.327. These results exceed those of structured-only, recurrent, gradient boosting, logistic regression, and text-only baselines.

The empirical pattern is consistent across several analyses. Text alone is useful but insufficient, structured time series remain the dominant signal source, and the best results arise when narrative context is fused with physiologic trajectories rather than treated as a substitute. Explicit modeling of masks and elapsed times improves robustness under irregular sampling, and site-adversarial training reduces hospital entanglement in the latent representation while improving external discrimination. Calibration is imperfect in the raw neural output but responds well to simple post-hoc temperature scaling. Under alert-budget simulation, the model concentrates a substantial fraction of future events within a relatively small high-risk subset of windows, which is favorable for ranked-review workflows.

The main implication is practical rather than rhetorical. In heterogeneous Chinese ICU environments, multimodal prediction appears most useful when coupled to transportability-aware training and workflow-conscious evaluation. The model should be understood as a continuous surveillance tool for prioritization and second-look review, not as an autonomous decision maker. Future work can extend this framework toward broader external validation, richer note-history encoding, outcome-specific heads, and prospective assessment of how ranked risk displays affect clinician behavior and patient trajectory [40].

References

[1] A. F. Muzaffar, S. Abdul-Massih, J. M. Stevenson, and S. Alvarez-Arango, "Use of the electronic health record for monitoring adverse drug reactions.," *Current allergy and asthma reports*, vol. 23, pp. 417–426, 5 2023.

[2] B. Sadanandan, "Multimodal deep learning for early prediction of patient deterioration in the icu: Integrating time-series ehr data with clinical notes," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 10, no. 7, pp. 1–21, 2025.

[3] K. Dharmayat, M. Woringer, N. Mastellos, D. Cole, J. Car, S. Ray, K. Khunti, A. Majeed, K. K. Ray, and S. R. K. Seshasai, "Investigation of cardiovascular health and risk factors among the diverse and contemporary population in london (the together study): Protocol for linking longitudinal medical records (preprint)," 12 2019.

[4] A. Herrmann-Werner, M. Holderried, T. Loda, N. P. Malek, S. Zipfel, and F. Holderried, "Navigating through electronic health records: Survey study on medical students' perspectives in general and with regard to a specific training," *JMIR medical informatics*, vol. 7, pp. e12648–, 11 2019.

[5] S. Visweswaran, M. J. Samayamuthu, M. I. Morris, G. M. Weber, D. MacFadden, P. Trevvett, J. G. Klann, V. S. Gainer, B. Benoit, S. N. Murphy, L. P. Patel, N. Mirkovic, Y. Borovskiy, R. D. Johnson, M. C. Wyatt, A. Y. Wang, R. W. Follett, N. Chau, W. Zhu, M. Abajian, A. Chuang, N. Bahroos, P. Reeder, D. Xie, J. Cai, E. R. Sendro, R. D. Toto, G. S. Firestein, L. M. Nadler, and S. E. Reis, "Development of a covid-19 application ontology for the act network," 4 2021.

[6] V. L. Tiase, W. Hull, M. M. McFarland, K. A. Sward, G. D. Fiol, C. J. Staes, C. R. Weir, and M. R. Cummins, "Patient-generated health data and electronic health record integration: a scoping review," *JAMIA open*, vol. 3, pp. 619–627, 12 2020.

[7] D. J. Slotwiner, "Electronic health records and cardiac implantable electronic devices: new paradigms and efficiencies," *Journal of interventional cardiac electrophysiology : an international journal of arrhythmias and pacing*, vol. 47, pp. 29–35, 9 2016.

[8] L. Colligan, H. W. W. Potts, C. T. Finn, and R. A. Sinkin, "Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record.," *International journal of medical informatics*, vol. 84, pp. 469–476, 3 2015.

[9] K. T. Kadakia, M. D. Howell, and K. DeSalvo, "Modernizing public health data systems: Lessons from the health information technology for economic and clinical health (hitech) act.," *JAMA*, vol. 326, pp. 385–386, 8 2021.

[10] M. F. Funk, "A survey of chiropractic intern experiences learning and using an electronic health record system.," *The Journal of chiropractic education*, vol. 32, pp. 145–151, 6 2018.

[11] K. Malm-Nicolaisen, A. J. Fagerlund, and R. Pedersen, "How do users of modern ehr perceive the usability, user resistance and productivity five years or more after implementation?," *Studies in health technology and informatics*, vol. 290, pp. 829–833, 6 2022.

[12] M. Y. Yang, G. H. Kwak, T. Pollard, L. A. Celi, and M. Ghassemi, "Evaluating the impact of social determinants on health prediction in the intensive care unit," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 333–350, ACM, 8 2023.

[13] M. A. Sarwar, T. Bashir, O. Shahzad, and A. Abbas, "Cloud-based architecture to implement electronic health record (ehr) system in pakistan," *IT Professional*, vol. 21, pp. 49–54, 5 2019.

[14] E. B. Devine, E. van Eaton, M. E. Zadworny, R. G. Symons, A. Devlin, D. Yanez, M. Yetisgen, K. R. Keyloun, D. Capurro, R. Alfonso-Cristancho, D. R.

Flum, and P. Tarczy-Hornoch, "Automating electronic clinical data capture for quality improvement and research: The certain validation project of real world evidence.," *EGEMS (Washington, DC)*, vol. 6, pp. 8–8, 5 2018.

[15] S. Bhartiya and D. Mehrotra, *Accessibility Issues in Interoperable Sharing of Electronic Health Records: Physician's Perspective*, pp. 199–218. Springer International Publishing, 3 2016.

[16] M. Tai-Seale, C. Wilson, C. Olson, M. Durbin, C. Morikawa, and H. Luft, "C3-1: Footprints in the sand: Tracking physician work efforts in primary care using access logs in an electronic health record," *Clinical Medicine & Research*, vol. 12, pp. 94–94, 9 2014.

[17] G. M. Belbin, S. Wenric, S. Cullina, B. S. Glicksberg, A. Moscati, G. L. Wojcik, R. Shemirani, N. D. Beckmann, A. Cohain, E. P. Sorokin, D. S. Park, J. L. Ambite, S. Ellis, A. Auton, C. G. Team, E. P. Bottinger, J. H. Cho, R. J. F. Loos, N. S. Abul-Husn, N. Zaitlen, C. R. Gignoux, and E. E. Kenny, "Towards a fine-scale population health monitoring system," 9 2019.

[18] Habibi-koolaee, R. Safdari, and H. Bouraghi, "Nurses readiness and electronic health records.," *Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH*, vol. 23, pp. 105–107, 4 2015.

[19] M. A. Sumon, T. A. Kwemebe, V. K. Melapu, and M. M. Hasan, "Emergency care patient prediction using electronic health records (ehr) data: An end-to-end machine learning pipeline," in *2023 10th International Conference on Future Internet of Things and Cloud (FiCloud)*, pp. 278–282, IEEE, 8 2023.

[20] J. Zhang, Y. Chen, S. Ashfaq, K. Bell, A. Calvitti, N. J. Farber, M. T. Gabuzda, B. Gray, L. Liu, S. Rick, R. L. Street, K. Zheng, D. Zuest, and Z. Agha, "Strategizing ehr use to achieve patient-centered care in exam rooms: a qualitative study on primary care providers.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 23, pp. 137–143, 11 2015.

[21] M. S. A. Malik, S. Sulaiman, and M. F. A. Malik, *Relational Factors of EHR Database in Visual Analytic Systems for Public Health Care Units*, pp. 161–172. Germany: Springer International Publishing, 9 2014.

[22] Z. Diao and F. Sun, "A deep-learning neural network approach for secure wireless communication in the surveillance of electronic health records," *Processes*, vol. 11, pp. 1329–1329, 4 2023.

[23] D. Hollar, "Progress along developmental tracks for electronic health records implementation in the united states," *Health research policy and systems*, vol. 7, pp. 3–3, 3 2009.

[24] T. Gizaw, M. Bogale, and T. Alemayehu, "Evaluation of the electronic health record system in maternal and child health centers of marie stopes international ethiopia," *Gates Open Research*, vol. 3, pp. 1655–, 11 2019.

[25] Y. Tarabichi, A. Frees, S. Honeywell, C. Huang, A. M. Naidech, J. H. Moore, and D. C. Kaelber, "The cosmos collaborative: A vendor-facilitated electronic health record data aggregation platform," *ACI open*, vol. 05, pp. e36–e46, 6 2021.

[26] S. M. Goodday, A. Kormilitzin, N. Vaci, Q. Liu, A. Cipriani, T. Smith, and A. J. Nevado-Holgado, "Maximizing the use of social and behavioural information from secondary care mental health electronic health records.," *Journal of biomedical informatics*, vol. 107, pp. 103429–, 5 2020.

[27] N. J. Giori, J. Radin, A. Callahan, J. A. Fries, E. Halilaj, C. Ré, S. L. Delp, N. H. Shah, and A. H. S. Harris, "Assessment of extractability and accuracy of electronic health record data for joint implant registries," *JAMA network open*, vol. 4, pp. e211728–, 3 2021.

[28] T. Hernandez-Boussard, D. W. Blayney, and J. D. Brooks, "Leveraging digital data to inform and improve quality cancer care.," *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, vol. 29, pp. 816–822, 4 2020.

[29] R. L. Altman, C.-T. Lin, and M. Earnest, "Problem-oriented documentation: design and widespread adoption of a novel toolkit in a commercial electronic health record.," *JAMIA open*, vol. 6, pp. ooad005–ooad005, 1 2023.

[30] A. Nishimwe, C. Ruranga, C. Musanabaganwa, R. Mugeni, M. Semakula, J. Nzabanita, I. Kabano, A. Uwimana, J. N. Utumatwishima, J. D. Kabakambira, A. Uwineza, L. Halvorsen, F. Descamps, J. Houghtaling, B. Burke, O. Bahati, C. Bizimana, S. Jansen, C. Twizere, K. Nkurikiyeyezu, F. Birungi, S. Nsanzimana, and M. Twagirumukiza, "Leveraging artificial intelligence and data science techniques in harmonizing, sharing, accessing and analyzing sars-cov-2/covid-19 data in rwanda (laisdar project): Study design and rationale," 3 2022.

[31] C. Violán, Q. Foguet-Boreu, E. Hermosilla-Pérez, J. M. Valderas, B. Bolíbar, M. Fàbregas-Escurriola, P. Brugulat-Guiteras, and M. Ángel Muñoz-Pérez, "Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity.," *BMC public health*, vol. 13, pp. 251–251, 3 2013.

[32] S. Cui, J. Wang, X. Gui, T. Wang, and F. Ma, "Automed: Automated medical risk predictive modeling on electronic health records.," *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, vol. 2022, pp. 948–953, 12 2022.

[33] Y. Zhao and K. Du, "A matching scheme from supply and demand sides of electronic health records based on blockchain," in *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, vol. 39, pp. 1089–1092, IEEE, 4 2022.

[34] C. C. Lee, Y. Kim, J. H. Choi, and E. Porter, "Does electronic health record systems implementation impact hospital efficiency, profitability, and quality?," *Journal of Applied Business and Economics*, vol. 24, 3 2022.

[35] K. M. VanLangen, K. G. Elder, M. Young, and M. Sohn, "Academic electronic health records as a vehicle to augment the assessment of patient care skills in the didactic pharmacy curriculum," *Currents in pharmacy teaching & learning*, vol. 12, pp. 1056–1061, 4 2020.

[36] S. T. Rosenbloom, W. W. Stead, J. C. Denny, D. A. Giuse, N. M. Lorenzi, S. H. Brown, and K. B. Johnson, "Generating clinical notes for electronic health record systems.," *Applied clinical informatics*, vol. 1, pp. 232–243, 1 2010.

[37] A. Khalifa, C. C. Mason, J. H. Garvin, M. S. Williams, G. D. Fiol, B. R. Jackson, S. B. Bleyl, and S. M. Huff, "A qualitative study of prevalent laboratory information systems and data communication patterns for genetic test reporting," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 23, pp. 2171–2177, 7 2021.

[38] E. Getzen, L. Ungar, D. Mowery, X. Jiang, and Q. Long, "Mining for equitable health: Assessing the impact of missing data in electronic health records," 5 2022.

[39] B. Kim, J. M. Chae, D.-S. Kim, H. Woo, C. Kim, H. B. Kim, H.-S. Kim, S. H. Park, S. J. Jeong, Y. Uh, S. V. Ahn, Y. S. Park, and J. Y. Choi, "1806. the gap in the amount of monthly antimicrobial use according to the data source: Electronic health record data vs. national health insurance claim data in korea," *Open Forum Infectious Diseases*, vol. 9, 12 2022.

[40] T. Adedeji, H. Fraser, and P. Scott, "Implementing electronic health records in primary care using the theory of change: A nigerian case study (preprint)," 9 2021.